**A MULTIDIMENSIONAL APPROACH TO THE IDENTIFICATION OF TEST BIAS: EXPLORATION OF THREE MULTIPLE-CHOICE SSC PAPERS IN PAKISTAN.**

*Syed Muhammad Fahad Latifi*

*Aga Khan University – Examination Board, Pakistan.*

*Centre for Research in Applied Measurement and Evaluation,*

*University of Alberta, Canada.*

*fahad.latifi@aku.edu*

*Dr. Thomas Christie*

*Aga Khan University – Examination Board, Pakistan.*

*thomas.christie@aku.edu*

**Abstract**

In Pakistan, the Secondary School Certificate (SSC) examination is a high-stakes national qualification at grade-ten. Consequently, fairness of SSC tests is an issue of concern among public and educational policy planners. During the test development process, substantive review is generally practised but the statistical aspect of test-fairness is an evolving topic among the examination authorities of the country.

This study is a first ever attempt, at the national level in Pakistan, to examine test-fairness using multidimensional Differential Item functioning (DIF) and Differential bundle functioning (DBF) procedures. The 2011 cohort of SSC examinees from Aga Khan University Examination Board (AKU-EB) is studied and multiple choice tests in three core subjects, English, Mathematics and Physics, are investigated for possible gender bias. All tests are based on National Curriculum (2006) guidelines from Ministry of Education (Curriculum Wing), Government of Pakistan.

The outcome from this study is expected to provide evidence of test fairness which will enhance test development practices in examination authorities of Pakistan and elsewhere.

**Keywords:** Validity; Multiple-Choice; Differential Item; Bias; SIBTEST

**INTRODUCTION**

*Background*

Testing programs are obligated to maintain fairness. Like any assessment setting, fairness in testing is an issue of concern for large-scale assessments in Pakistan. The statistical aspect of test-fairness is an evolving topic among the examination authorities in the country. The present study is the first attempt to analyze the test fairness using two related statistical procedures: Differential Item functioning (DIF) and Differential Bundle functioning (DBF). The results from this study are expected to give insights into the test fairness in the country. Three core subjects from Annual Secondary School Certificate-examination (Grade-10), administered by the Aga Khan University Examination Board in 2011, are evaluated for fairness using DIF and DBF analyses.

DIF occurs when examinees with the same ability but in different groups have a different probability of answering the test item correctly. DBF is a concept build upon DIF, in which a subset of items within a test is organized to form a bundle of items. These subtests or item-bundles are then analyzed for any potential differential performance among the groups after controlling for ability. Certain organizing principle needs to be followed while creating a bundle of items (Gierl, Bisanz, Bisanz, Boughton, & Khaliq, 2001).

The present study examines the DIF due to gender differences for three core subjects namely English, Mathematics and Physics, which are part of the Annual Secondary School Certificate (SSC) examination. The examinations were administered by the Aga Khan University Examination Board (AKU-EB) during May-2011 across Pakistan, to students to see if they met the qualification for certificates at two levels: Secondary School Certificate (SSC, Grade-10) and Higher Secondary School Certificate (HSSC, Grade-12). The AKU-EB is an autonomous Federal Board of Intermediate and Secondary Education established in 2003 with its first practice examination administered in 2006 followed by full scale examination administered in May 2007 and afterwards. The outcome from this study is expected to provide insights about the quality of tests from a fairness perspective, which can be used to enhance item and test development practices at AKU-EB, and will be beneficial to other examination authorities in the country and elsewhere.

*Test Development Process at AKU-EB*

 At AKU-EB all examination materials are developed by content experts and every item under goes a comprehensive review process including a sensitivity review. Items are developed by teachers and panel experts nominated to serve on item writing and review committees. These committees are formed by Curriculum and Exam Development (C&ED) unit of AKU-EB. Items are field tested after first internal review and classical item analysis indices from field testing are used to screen and revise the items. For each test, the final review is conducted by a panel of experts, comprising of two senior subject teachers from affiliated schools and an internal C&ED subject expert. The items are evaluated for content and curriculum coverage, item appropriateness, grammatical accuracies, and sensitivity analysis for any possible biases. The final test items are then selected by C&ED subject experts, who prepare two parallel test-forms using common set of test specifications across years. Of these two test-forms, one is chosen by two senior members from C&ED for actual exam administration; the other test-form is secured for addressing any unforeseen exam emergency. The detailed procedure of test development process has been documented in AKU-EB’s Standard Operation procedure (revised edition, 2010-2011).

*Purpose and Scope*

The present study used test data from SSC (Grade-10) May 2011 administration. The English, Mathematics and Physics examinations were chosen based on their relative importance of the results obtained from these tests in determining career path of the examinees. Each subject has two paper components: Multiple Choice Question (MCQ, paper-I) and Constructed Response Question (CRQ, paper-II). For the purpose of this study only MCQ portion of the exam is studied. The MCQ test for English is composed of 25 items; the Mathematics test includes 30 items and Physics test contains 25 items. All items were dichotomously scored (content wise details is presented in Appendix-A1). All tests are based on National Curriculum (2006) guidelines from Ministry of Education (Curriculum Wing), Government of Pakistan, which describe the Competencies, Standards and Benchmarks for SSC/HSSC assessments.

In Pakistan, SSC (Grade-10) is considered extremely high stakes and to qualify as SSC each examinee has to pass eights subjects. In the case that an examinee fails, two years of academic efforts will go in vain and the chances to continue academics at college and university level will be minimized. Consequently, it becomes important that the test and test-items should not favor one group of examinees over another. This study is expected to evaluate and verify the test fairness practices which are built into the test development framework of AKU-EB, the outcome from this study, may highlight anomaly or, may verify the fairness and accuracy of test development framework at AKU-EB.

**METHOD**

*DIF Detection Procedure*

For the purpose of this study, exploratory DIF analysis was first conducted. The first step in any such analysis is the selection of appropriate DIF detection procedure. Although many DIF methods are available, a relatively small number of these methods are “preferred” based on their theoretical and empirical strengths (Gierl, Gotzmann & Boughton, 2004). The Simultaneous Item Bias Test or SIBTEST (Shealy & Stout, 1993) is one of the preferred methods. SIBTEST is used to compute a weighted mean difference between the reference and focal groups, and this difference is tested statistically. The means in this procedure are adjusted to correct for any differences in the ability distributions of the reference and focal groups using a regression correction procedure described by Shealy and Stout that, in effect, creates a matching subtest (common measure) free from statistical bias. Using a simulation study the authors of the SIBTEST established that the method is only marginally affected by a large number of DIF items in the matching subtest. Recently, Gierl et al. (2004) evaluated SIBTEST, using multiple simulated datasets, for its accuracy in detecting DIF items when the percentage of DIF items on the test is large, and the findings were consistent and accurate.

*Multidimensional DIF Analysis Framework*

The Multidimensional DIF analysis framework employs the concepts of primary and secondary dimensions to explain the DIF in an item. Dimension refers to a substantive characteristic of an item that can affect the probability of a correct response on the item. Each item in a test is intended to measure the main construct called primary dimension. DIF items measure at least one dimension in addition to the primary dimension (Ackerman, 1992; Roussos & Stout, 1996a; Shealy & Stout, 1993; as cited in Boughton, Gierl & Khaliq, 2000). This additional dimension, which produces DIF, is referred to as the secondary dimension. A secondary dimension is named as auxiliary-dimension if intentionally assessed or a nuisance-dimension if no certain reason is established for its existence. DIF caused by auxiliary dimensions is benign whereas DIF caused by nuisance dimensions is adverse and thus reflects bias for examinees in either the reference or the focal group. The reference group is the majority group or the group to whom the focal group is compared. The focal group is the minority group or the particular group of interest in the DIF analysis.

*SIBTEST DIF Index (Beta-Uni)*

To identify items that function differentially for Male and Females, SIBTEST was used, with males as reference group and females as focal group. SIBTEST provides a measure of effect size, Beta‑uni (UNI), with overall statistical test for each test item or bundle. In this statistical approach, the complete latent space is viewed as multidimensional, (*θ,η*), where *θ* is the primary dimension and *η* is the secondary dimension. The statistical hypothesis tested by SIBTEST is H0: UNI = 0 *vs.* H1: UNI ≠ 0.

UNI estimates the magnitude of DIF and provides measure of effect size. To operationalize this statistical approach, items on the test are divided into the suspect subtest and the matching or valid subtest. The suspect subtest contains the items or bundle of items believed to measure both primary and secondary dimensions whereas the matching subtest contains the items believed to measure only the primary dimension. The matching subtest places the reference and focal group examinees into subgroups at each score level so their performances on items from the suspect subtest can be compared. To estimate UNI, the weighted mean difference between the reference and focal groups on the suspect subtest item or bundle across the *K* subgroups is calculated by The test statistic in SIBTEST is formed by dividing the estimate of UNI by an estimate of its standard error, called SIB. Where is calculated by . A statistically significant value of the UNI that is positive indicates DIF against the focal group whereas a negative value indicates DIF against the reference group. Thus the inspection of sign of the SIBTEST statistics can then be used to help determine against which group the DIF (or DBF) is likely acting.

To classifying the SIBTEST effect sizes (UNI), guidelines recommended by Dorans (1989) for interpreting standardized *p*-values were used. Since both the standardization and SIBTEST methods are essentially measuring the same construct (i.e., the total test score), using Doran’s criteria in the SIBTEST context seemed reasonable (Puhan, Boughton, & Kim, 2007). Therefore, absolute values of the UNI statistic less than 0.050 indicate negligible DIF (level-A), between 0.050 and 0.099 indicate moderate DIF (level-B) and 0.100 and above indicate large DIF (level-C). Based on these guidelines, items are classified as “A” (negligible DIF), “B” (moderate DIF) or “C” (large DIF). Also, to determine statistical significance, an alpha level of 0.05 was used.

*Bundling of Test Items*

Douglas, Roussos, and Stout (1996) suggested that the DIF at the item level may be level-A DIF items or may go statistically undetected in the single item DIF detection approach but can be detected using the bundle approach. DBF analysis requires items to be organized using certain organizing principles. To identify the dimensions of tests, four organizing principles have been suggested in literature (Gierl, 2005) and can be used to organized test items. They are: 1) Test Specification, 2) Content Analysis, 3) Psychological Analysis, and 4) Empirical Analysis. For the purpose of the present study, the DBF analysis will be based on creating bundles using the test specification for the three examinations considered. The test specifications for each examination are presented in Appendix A-1.

**RESULTS**

*Psychometric Characteristics of Data*

To study psychometric characteristics the classical indices were computed using BILOG-MG. The outcomes are summarized in Table 1. As shown in Table 1, the number of females outnumbered the number of males by approximately 100 across the three subjects (103, 118 and 98 respectively). While the mean test scores of the females were greater than the mean scores for males, the unpaired t-test is computed for each subject and only English is found with significant (*p* < 0.05) mean difference. The values of Cohen’s *d* statistic – 0.210, 0.014 and 0.021 – were found weak which suggested that the gender grouping is comparable.

Likewise, the difference between the standard deviations, skewness and kurtosis of the distributions of scores for the females and the males are small. Further, the mean item difficulty and mean discrimination values are also comparable for both groups. The reliability of Mathematics examination is the highest (0.89 for both male and females) while the reliabilities of the English and Physics examination were more comparable (0.79 and 0.81 respectively), with the higher value for Cronbach`s α for Mathematics reflects the higher discrimination for the Mathematics items. This is likely due to the more structure nature of Mathematics compared to English and Physics.

The results presented in Table 1 indicate that the test developers were successful in minimizing gender difference at the test-score level; there was no difference in mean performance between focal and reference groups for the English, Mathematics and Physics examinations. Further, the similarity of the other statistics in each subject suggests that the AKU-EB guidelines for item and test construction are being followed.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table 1 |  |  |  |  |
| *Psychometric Characteristics for the SSC English, Mathematics and Physics.* |
|  |  |  |  |  |  |  |  |
|  |  | **English** | **Mathematics** | **Physics** |
| **Characteristics** | Male | Female | Male | Female | Male | Female |
| No of Examinees | 982 | 1085 | 819 | 937 | 818 | 916 |
| No. of Items | 25 | 25 | 30 | 30 | 25 | 25 |
| Mean | 14.14 | 15.03 | 19.21 | 19.29 | 15.41 | 15.49 |
| SD | 4.27 | 4.21 | 6.04 | 5.88 | 4.13 | 4.07 |
| Skewness | -0.08 | -0.24 | -0.30 | -0.31 | -0.41 | -0.39 |
| Kurtosis | -0.45 | -0.40 | -0.65 | -0.67 | -0.55 | -0.48 |
| Mean Item Difficulty | 0.57 | 0.60 | 0.64 | 0.64 | 0.62 | 0.62 |
| SD Item Difficulty | 0.20 | 0.20 | 0.15 | 0.16 | 0.16 | 0.16 |
| Mean Item Discriminationa | 0.28 | 0.28 | 0.38 | 0.37 | 0.26 | 0.26 |
| SD Item Discrimination | 0.09 | 0.10 | 0.09 | 0.09 | 0.15 | 0.14 |
| Internal Consistencyb | 0.79 | 0.79 | 0.89 | 0.89 | 0.81 | 0.81 |
| aItem-to-Total Pearson (point-biserial)bCronbach`s Alpha Coefficient |  |  |  |  |  |

*Exploratory DIF – Phase-one*

After establishing the test reliability and comparability at the group level, the phase one exploratory DIF analysis was initiated by estimating UNI for each test item in all three subjects. For this phase, each test item is considered as a suspect subtest while the remaining items were considered as the matching subtest. This process was then repeated for all the test items, one by one. The sign of UNI determines the direction of DIF, positive UNI value suggest that the item favor males whereas negative value favor female. The results from phase one are displayed in Figure 1.

**

*Figure 1.* Phase one: Gender differences for items from SSC English, Mathematics and Physics May 2011 Examinations.

Figure 1 represents the DIF estimate for items for the SSC May 2011 English, Mathematics and Physic examinations. Examination of the figure reveals that one item in each subject examination had large DIF (level-C). Whereas Item 14 in the English test favored Males and Item 11 in the Mathematics examination favored Males, Item 4 in the Physics examination favored Females. The other notable attribute is the spread of UNI values, which is not skewed in any favoring direction. Also there are some level-B DIF items which are by definition producing moderate DIF and they are distributed evenly between male and females, but they are of little concern at this stage. As moderate DIF items would be more difficult to flag than large DIF items and most DIF is classified as moderate in large-scale testing programs (Linn, 1993; as cited in Gierl, Gotzmann & Boughton, 2004).

Statistically significant UNI value: for item 14 in English is 0.100, for item 11 in Mathematics is 0.105, and for item 4 in Physics is -0.150. A substantive analysis is required to interpret for all of these flagged items. For example, for Physics, the hypothesis may be developed that since females are more interested in medical careers than males, they are more proficient than males when dealing with the use of ultrasound. The details about these level-C questions can be inspected in Appendix A-2.

*Creating Bundle for DBF Analysis*

The bundles of items for each subject examination were formed in terms of the Table of Specification for the subarea (see Appendix A-1). The distributions of DIF for each subarea are displayed in Figure 2. As shown, there are two subareas in English: *Listening Skills* and *Reading Skills*; three subareas in Mathematics: *Coordinate Geometry, Trigonometry and Theorems*, *Fraction, Functions and Algebraic Manipulation* and *Linear and Quadratic Equations, Inequalities and Graphs*; three subareas in Physics: *Electronics, ICT and Radioactivity*, *Electrostatics, Current and Magnetism* and *Waves, Sound and Optics*. At the bottom of the graph, Beta-Uni values, for the bundle items in each subarea are provided.

**

 *Figure 2.* Phase Two: Gender difference for items from SSC English, Mathematics and Physics May-2011 Examinations, organized into bundles using the test specification.

As shown in Figure 2, content area *Reading Skills* carries the level-C DIF item that favored males in English and males generally performing better on the Reading Skills items, *Equation, Inequalities and Graph* in Mathematics carries a level-C item and it favoring males. Physics is generally balanced, with one level-C item (UNI < -0.15) in *Waves, Sound and Optics* favoring females, also the spread of UNI value is not skewed in any favoring direction. The substantive analysis may reveal whether this DIF is due to impact or bias.

*Analysis of Bundle Functioning*

After considering the test specification chart and pattern from content-wise UNI values from Figure 2, different item bundles were tested. As shown at the bottom of the Figure 2, the DBF for three of the seven bundles was found to be statistically significant (*p* < 0.05).The DBF favored females in the subarea *Listening Skills*, favored males in subarea *Fraction, Functions and Algebraic Manipulation*, and males in the subarea of *Waves, Sounds and Optics*. Interestingly, the significant bundles for English and Mathematics tests do not carries large (level-C) DIF item. Interpreting DBF requires caution as there is no guideline that exists for classifying the value of UNI for bundle of items and as a result there is no agreement on how to distinguish statistical from practical significance in DBF research (e.g. Gierl, Bisanz, Bisanz & Boughton, 2002). To overcome this limitation, Gierl et al. (2002), for example, suggested supplementing a DBF analysis using DIMTEST to evaluate the dimensional structure of the data which is assumed in the SIBTEST. However, this was not part of the present study where the bundles were formed using Test Specifications which are meaningful to test developers and agencies. The Validity of the DBF finding can further analyzed using substantive analyses, in which content experts may be engaged to elaborate whether these findings are benign or adverse, or by replicating the study using the May 2010 SSC English, Mathematics, and Physics examinations.

**DISCUSSION**

Certificate, diploma and other high-stakes examinations such as the SSC will continue to be used for making critical decision about the examinees and thus will affect an individual`s career path. Consequently it is imperative that test items and the test must be free from any fairness objection. The findings from this study suggest that the three examinations are free from any fairness objection. The exploratory data analysis was initiated by evaluating the comparability between males and female in three subjects (namely English, Mathematics and Physics). First, the classical test indices were computed including internal consistency measure for evaluating the reliability of tests. Second the score distribution was evaluated for comparability using effect size measure (i.e. Cohen`s *d* statistics) and groups were found comparable on each subject except in English where groups are significantly different on their mean score, as shown in Table 1. However, the effect size measure suggested the difference between males and females were small and weak, suggesting that at the test level males and females were essentially the same and that neither was favored.

The DIF analysis at the item revealed that one item in each subject area possessed large (or level-C) DIF, with two of these items favoring males and one favoring females. Given the total number of items with less C-level DIF items, no noticeable DIF was found, an indication that the test development practices remained successful in ensuring the fairness, as individual test items were largely found free of gender bias.

To further establish the purity of findings, the DBF analysis was conducted where the bundle of items were created using the Table of Specification for each examination. Of the seven suspect bundles created, only three were found to be statistically significant. The DBF reveals that the females are performing better than males on items requiring *Listening Skills*. Alternatively, males are performing better than females on items requiring *Fraction, Functions & Algebraic Manipulation*; and for *Waves, Sounds and Optics* (see bottom of Figure 2).

The affect of gender in differential performance on content and cognitive skills was studied by Gierl, Bisanz, Bisanz, & Boughton (2003) and they concluded that males perform better than females on items that require significant spatial processing and that females perform better than males on items requiring memorization ( or memory recall). Since *Listening skills* were tested using a recorded dialogue, which was played at the exam hall using CD/ Cassette players and examinees are expected to recall dialogue and answer the test items, the finding from the present study is consistent with what Gierl et al.(2003) found. Also the spatial aspect of *Fraction, Functions & Algebraic manipulation* and *Waves, Sounds and Optic* from Mathematics and Physics may be favoring males. These finding are also in accordance with the finding from Gierl et al.(2003), However similar supportive arguments are not present for the four bundles that were not significant.

**CONCLUSION**

Taken together, the results of this study indicates that there were only three items with Level-C DIF in the AKU-EB`s SSC May 2011 English, Mathematics and Physics examinations. Further, while three of the seven bundles of items were found to significantly favor males (two bundles) or females (one bundle), these bundles appeared in each subject area and although significant, did not seem to be that large in comparison to the non-significant statistics for the other four bundles. Taken together, the present study suggests that the small amount of DIF found does not constitute threats to the overall validity of the interpretation of the examinees’ test scores on SSC examination.

**REFERENCE**

Boughton, K., Gierl, M. J., & Khaliq, S. (2000, May). Differential bundle functioning on mathematics and science achievement tests. Paper presented at the annual meeting of the *Canadian Society for Studies in Education,* Edmonton, Alberta, Canada.

Dorans, N.J. (1989). Two new approaches to assessing differential item functioning: Standardization and the Mantel-Haenszel method. *Applied Measurement in Education*, 2(3), 217–233.

Douglas, J. A., Roussos, L. A., & Stout, W. (1996). Item-bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning. *Journal of Educational Measurement*, 33, 465-484.

Gierl, M.J., (2005). Using Dimensionality-Based DIF Analyses to identify and Interpret Constructs That Elicit Group Differences, *Educational Measurement: Issues and Practice*, 24, 3-14

Gierl, M. J., Bisanz, J., Bisanz, G., & Boughton, K. (2003). Identifying content and cognitive skills that produce gender differences in mathematics: A demonstration of the DIF analysis paradigm. *Journal of Educational Measurement*, 40, 281-306.

Gierl, M.J., Bisanz, J., Bisanz, G., Boughton, K., & Khaliq, S. (2001). Illustrating the utility of differential bundle functioning analysis to identify and interpret group differences on achievement tests. *Educational Measurement: Issues and Practice*, 20, 26-36.

Gierl, M.J., Bisanz, J., & Li, Y., (2004). Using the Multidimensionality-Based DIF Analysis Paradigm to Study Cognitive Skills that Elicit Group Differences: A Critique. Paper Presented at the Annual Meeting of the *National Council on Measurement in Education* .

Gierl, M.J., Gotzmann, A., & Boughton, K., (2004): Performance of SIBTEST When the Percentage of DIF Items is Large, *Applied Measurement in Education*, 17:3, 241-264

Puhan, P., Boughton, K., & Kim, S. (2007). Examining Differences in Examinee Performance in Paper and Pencil and Computerized Testing. *Journal of Technology, Learning, and Assessment, 6*(3). Retrieved [April 2011] from <http://www.jtla.org>.

Shealy, R., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika*, 58, 159-194.

**APPENDIX**

|  |
| --- |
| ***Appendix A-1*** |
| *Test Specification for MCQ English, Mathematics and Physics examination May-2011.* |
|   |   |   |   |
| **Subject** | **Subject Content Area** | **Items no. on test form** | **No of Items.** |
| *English* |  |  |  |
|  | Listening Skills | 1 to 12 | 12 |
|  | Reading Skills | 13 to 25 | 13 |
| *Mathematics* |  |  |  |
|  | Coordinate Geometry, Trigonometry and Theorems. | 15 to 30 | 16 |
|  | Fraction, Functions and Algebraic Manipulation | 2 to 5 | 4 |
|  | Linear & Quadratic Equations, Inequalities and Graphs | 1,6 to 14 | 10 |
| *Physics* |  |  |  |
|  | Electronics, ICT and Radioactivity | 17 to 25 | 9 |
|  | Electrostatics, Current and Magnetism | 9 to 16 | 8 |
|   | Waves, Sound and Optics | 1 to 8 | 8 |

